Distribution of Sample Means and a Confidence Interval for the Population Mean

Shannon Pileggi

STAT 217

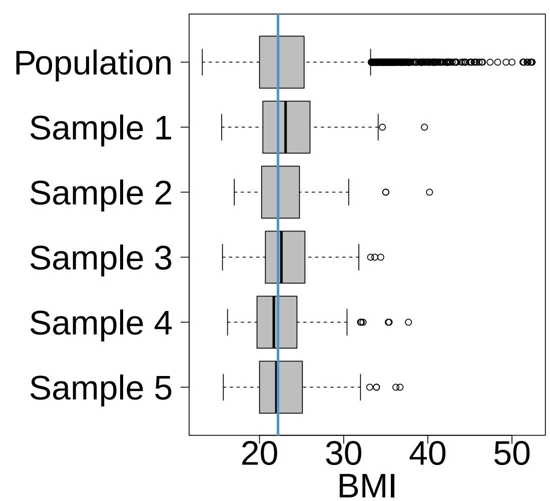
OUTLINE

The Data

From the CDC's 2013 Youth Risk Behavior Surveillance System

	gender	height_m	weight_kg	bmi	carried_weapon	bullied	days_drink
1	female	1.73	84.37	28.2	yes	no	30
2	female	1.6	55.79	21.8	no	yes	1
3	female	1.5	46.72	20.8	no	yes	0
4	female	1.57	67.13	27.2	no	yes	0
5	female	1.68	69.85	24.7	no	no	0
6	female	1.65	66.68	24.5	no	no	1
7	male	1.85	74.39	21.7	no	no	0
8	male	1.78	70.31	22.2	yes	no	0
9	male	1.73	73.48	24.6	no	yes	0
10	male	1.83	67.59	20.2	no	no	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
8482	male	1.73	68.95	23	no	no	0

The idea



- the entire data set represents a population
- each sample is of size $n = 100$
- the blue line is the median BMI in the entire data set

Population distribution of days_drink

For this exercise, consider the 8,482 observations from the YRBSS data set to be the *entire* population of interest. Now let's describe the **population distribution** of `days_drink`.

- Shape of the population distribution:
- Mean of the population distribution:
- Standard deviation of the population distribution:

Example data distributions from days_drink

Now let's take three random samples of size $n = 10$ from the population distribution of `days_drink`. Each random sample represents a **data distribution**.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	shape	$\bar{x}$	s
Sample 1													
Sample 2													
Sample 3													

Many samples from days_drink

Let's repeat the process and take 1000 random samples of size $n = 10$ from the population distribution of `days_drink`.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	$\bar{x}$	s
Sample1	1	0	9	0	1	3	0	3	0	0	1.7	2.83
Sample2	0	0	0	0	0	0	1	1	3	0	0.5	0.97
Sample3	4	1	0	0	0	0	0	1	2	0	0.8	1.32
Sample4	0	0	0	0	0	0	2	0	0	1	0.3	0.67
Sample5	0	1	0	0	0	0	0	0	1	0	0.2	0.42
Sample6	0	0	0	3	0	0	0	1	6	2	1.2	1.99
Sample7	0	0	30	0	0	0	2	0	0	5	3.7	9.38
Sample8	0	0	9	1	1	0	0	0	8	0	1.9	3.51
Sample9	1	0	4	0	0	4	0	4	1	0	1.4	1.84
Sample10	0	0	0	1	0	0	0	4	0	0	0.5	1.27
Sample11	0	0	0	0	0	5	0	0	0	0	0.5	1.58
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Sample1000	0	0	13	0	0	3	0	0	0	15	3.1	5.84

Clicker

What do you think will be the shape of the distribution of the 1000 sample means?

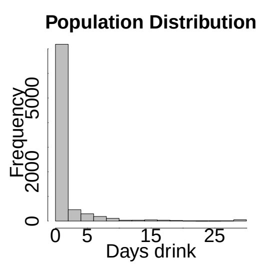
1. bell-shaped
2. left-skewed
3. right-skewed
4. uniform

Simulated sampling distribution, example 1

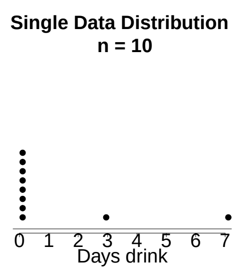
The collection of the sample means from the 1000 samples of size $n = 10$ represents a simulated **sampling distribution** of the sample mean.

- Shape of the sampling distribution:
- Mean of the sampling distribution:
- Standard deviation of the sampling distribution:

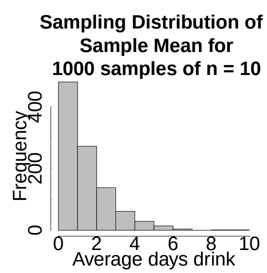
Re-cap, example 1



mean $= 1.45$
sd $= 3.78$



mean $= 1.00$
sd $= 2.31$



mean $= 1.39$
sd $= 1.19$

Group Exercise

What do you think will happen to the distribution of sample means if we increase the sample size for each individual sample from $n = 10$ to $n = 200$? (The number of samples will stay the same at 1000.)

The shape will be ____________, the mean will ____________, the standard deviation will ____________.

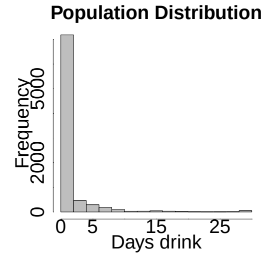
1. shape: right-skewed, left-skewed, approximately normal
2. mean: increase, decrease, remain the same
3. standard deviation: increase, decrease, remain the same

Simulated sampling distribution, example 2

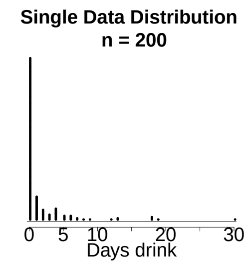
The collection of the sample means from the 1000 samples of size $n = 200$ represents a simulated **sampling distribution** of the sample mean.

- Shape of the sampling distribution:
- Mean of the sampling distribution:
- Standard deviation of the sampling distribution:

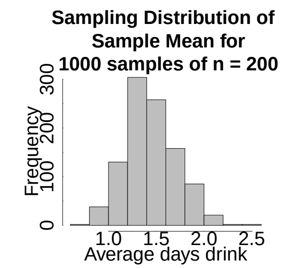
Re-cap, example 2



mean = 1.45
sd = 3.78



mean = 1.55
sd = 3.87



mean = 1.45
sd = 0.23

Summary

Feature	Example 1 ($n = 10$)	Example 2 ($n = 200$)
Observed in simulation		
Shape		
Mean		
Std Dev		
According to theory		
Shape		
Mean		
Std Dev		

Distribution of a Sample Means

OR: the sampling distribution of the sample mean

When sampling from a population with mean μ and standard deviation σ the **sampling distribution** of the **sample mean** has

$$\text{mean} = \mu \text{ and standard deviation} = \frac{\sigma}{\sqrt{n}}$$

Saying the same thing, but with more notation:

$$\text{mean}(\bar{x}) = \mu,\ \text{sd}(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Distribution of a Sample Means

OR: the sampling distribution of the sample mean

When the population is normally distributed, then the distribution of sample means is **approximately normal** regardless of your sample size n.

That is,

- shape = normal
- mean = μ
- standard deviation = $\frac{\sigma}{\sqrt{n}}$

for a normally distributed population, regardless of n.

Sampling Distribution of a Sample Mean

Central Limit Theorem

Regardless of the shape of the underlying population distribution, as the sample size n increases the distribution of sample means becomes approximately normal distribution.

That is, for large n,

- shape = normal
- mean = μ
- standard deviation = $\frac{\sigma}{\sqrt{n}}$

regardless of the shape of the underlying population distribution.

The distribution of sample means is usually close to bell shape when the sample size n is at least 30.

Sampling Distribution of a Sample Mean

Population Distribution	**Sample Size**	**Distribution of Sample Means** *Mean, SD*	*Shape*
Normal	Large	mean $= \mu$, sd $= \frac{\sigma}{\sqrt{n}}$	normal
Normal	Small	mean $= \mu$, sd $= \frac{\sigma}{\sqrt{n}}$	normal
Other	Large	mean $= \mu$, sd $= \frac{\sigma}{\sqrt{n}}$	normal
Other	Small	mean $= \mu$, sd $= \frac{\sigma}{\sqrt{n}}$	$\bar{x}$ not normal

Assume a simple random sample is used to gather data. Then, as you collect more data (n increases), which of the following is false?

1. You expect a histogram of the data distribution to look more and more like a normal distribution.
2. You expect the data distribution to resemble more closely the population distribution.
3. The sample mean tends to get closer to the population mean.
4. By the central limit theorem, the sampling distribution tends to take on more of a bell shape.

Which of the following affects the variability in the sampling distribution of the sample mean? Select all that apply

1. the population mean
2. the population standard deviation
3. the sample size
4. the number of samples collected

The number of calories in a cheeseburger is normally distributed with a mean of 500 and a standard deviation of 100. We take a random sample of 10 cheeseburgers.

Can we assume that the sampling distribution of the sample mean calories is approximately normally distributed?

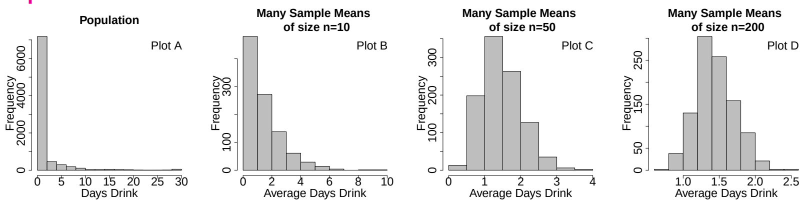
1. Yes, because the underlying population is normal
2. No, because n is small
3. Yes, because $np > 10$ and $n(1-p) > 10$
4. No, because one of $np > 10$ and $n(1-p) > 10$ is not satisfied
5. Not enough information to determine.

Example

The number of calories in a cheeseburger is normally distributed with a mean of 500 and a standard deviation of 100. Sketch:

1. the population distribution of calories in a cheeseburger
2. the distribution of sample mean calories for samples of size $n = 10$ cheeseburgers

Group Exercise



Would it be surprising to see (and why?)

1. A teenager drink more than 4 days
2. The average number of days drink of 10 teenagers to be greater than 4
3. The average number of days drink of 50 teenagers to be greater than 4
4. The average number of days drink of 200 teenagers to be greater than 4

Group Exercise

Which of these plots do you think the 68-95-99.7 rule applies to? Mark all that apply.

1. Plot A
2. Plot B
3. Plot C
4. Plot D

The idea

- Before, we discussed the distribution of means when we take many samples from a population.
- In practice, we only take one sample from a population! How do use one sample from a population to estimate the mean of a population?
- We use our estimates from our one sample $(\bar{x}, s)$ to construct a confidence interval.
- This method relies on the properties of the normal distribution, which is why we need to assess if our sampling distribution is normal or not.

CI for a population mean

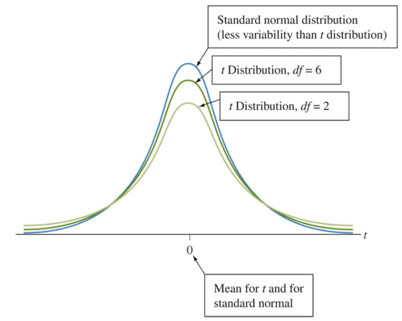
$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

point estimate $\pm$ *critical value* $\times$ *standard error*

point estimate $\pm$ *margin of error*

- the **point estimate** is your best guess of a population parameter $\rightarrow \bar{x}$
- the **critical value** establishes your degree of confidence for that interval $\rightarrow$ use t^*, $df = n - 1$
- the **standard error** allows for uncertainty in that point estimate $\rightarrow \frac{s}{\sqrt{n}}$
- the **margin of error** is the (critical value $\times$ standard error), and is everything after the $\pm$ $\rightarrow t^* \frac{s}{\sqrt{n}}$

The *t*-distribution



- bell-shaped and symmetric about 0
- like the normal distribution, but "fatter"
- characterized by the *degrees of freedom* (df).
- $df = n - 1$, determines how "fat" the *t*-distribution is

Critical values for 95% confidence level:

$t^*_{df=5}$	$t^*_{df=20}$	$t^*_{df=40}$	$t^*_{df=500}$	z^*
2.57	2.09	2.02	1.96	1.96

Conditions required for a CI for μ

1. The observations are independent.
2. The population distribution is normal *or* we have a 'large' sample size ($n \geq 30$).

Steps to constructing a confidence interval for a population mean.

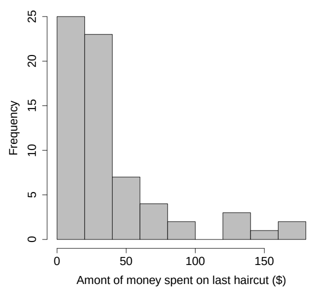
1. Check your conditions.
2. Identify t^* for your specified level of confidence ($df = n - 1$).

$t^*_{df=5}$	$t^*_{df=20}$	$t^*_{df=40}$	$t^*_{df=500}$	z^*
2.57	2.09	2.02	1.96	1.96

3. Calculate the interval: $\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$

Group Exercise

Here we have sample data from 50 Cal Poly students regarding the amount spent on their last hair cut. Use this to estimate the population average amount of money spent on haircuts by all Cal Poly students with a 95% confidence interval.



$n = 50$
$\bar{x} = 40$
$s = 25$

A 95% CI for average amount spent on a haircut is (32.90,47.09).

Which of the following provide correct interpretations of this confidence interval? Mark all that apply.

1. With 95% confidence, the average amount spent on a haircut by Cal Poly students in this sample is between 32.90 and 47.09.
2. With 95% confidence, Cal Poly students on average spend between 32.90 and 47.09 on a haircut.
3. A randomly chosen Cal Poly student has a 0.95 probability of spending between 32.90 and 47.09 on a haircut.
4. 95% of Cal Poly students spend between 32.90 and 47.09 on a haircut.

Elements of an interpretation of a confidence interval

1. State the confidence level
2. Refer to the population
3. State the parameter being estimated
4. Utilize context
5. Include a range of values

At the __1__ % confidence level, we estimate that the __2__ __3__ of __4__ is in the interval __5__.

Group Exercise

$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

What factors affect the width of the CI?

1.

2.

3.